Traffic Accidents Analytics in UK Urban Areas using *k*-means Clustering for Geospatial Mapping

Christopher Sinclair, *Student Member, IEEE*
Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn Campus, Cornwall TR10 9FE, United Kingdom.
Email: chrissinclair@ieee.org, cs797@exeter.ac.uk

Saptarshi Das, *Member, IEEE*
Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn Campus, Cornwall TR10 9FE, United Kingdom.
Email: saptarshi.das@ieee.org, s.das3@exeter.ac.uk

***Abstract*—The goal of this paper is to use the unsupervised machine learning method in road accident analytics, especially using *k*-means clustering to identify patterns and understand the relationships between variables recorded by the UK police department. These include features like number of casualties, number of vehicles, age of vehicle and age bracket of the driver. We aim to describe clusters of accidents based on similarity measures in the features and identify what separates each one.**

Keywords—clustering, road accident, geospatial analytics

I. Introduction

Despite the improvements in transport technologies, road accidents in the UK are very common and cause not only significant damage to those directly affected but also cause significant inconvenience to the public [1]. However, due to the high complexity of each different accident based on driver's fault, weather, road condition, specific location etc. it is almost impossible to build a mechanistic or first principle model (as opposed to a statistical model) to get deeper insight into how accidents happen and how we can prevent them [2]. The only feasible solution to understand the mechanism of accidents is perhaps analysing historic events and studying their patterns using different statistical modelling and machine learning methods. Machine learning methods have previously been used in the context of vehicle accident analytics such as various data mining approaches including classification and association rule mining [3],[4],[5], hybrid clustering regression approach [6], latent class clustering and Bayesian networks [7] or spatio-temporal clustering [8]. The *k*-means clustering has also been used successfully in profiling vehicle accident hotspots [9], identifying traffic congestion [10], modelling vehicle trajectories at cross roads along with fuzzy *c*-means clustering [11] for examining patterns of vehicle crashes in before-after analysis.

In the interest of facilitating to answer these questions, the UK Government's Department of Transport has released large datasets including all road accidents between 2005-2017 [12]. The number of car crashes per year is gradually on the decline in the UK. This is likely due to the technological progress of cars themselves rather than other external factor such as improving policing methods as numbers of road police fell by 27% between 2010-2014, while accidents per month maintained a downward trend. Besides road policing focuses primarily on reducing fatal and serious accidents, usually the result of driving while under the influence of alcohol or drugs. This downward trend shows that vehicle accidents are very much a solvable problem, and that continued efforts in this field will gradually bring about life saving changes in the road systems of both developed and developing countries. With improved vehicle designs, this will become especially true as driverless vehicles using advanced data science technologies transition into the normal life. On contrary to these examples, we here focus mainly on higher and reduced dimensional clustering of road accident data in UK and map the variables on global and local geospatial scales for better understanding of the relationships of the features for such accidents.

II. Materials and Methods

A. Description of the Dataset

The dataset under consideration is obtained from Kaggle [12], which is a compendium of UK police traffic reports detailing accident information ranging from the year 2005 to 2017 and vehicle information ranging from 2004 to 2016. It covers a wide range of 34 variables, though as discussed in the 'variables' section some, if not most are not useable within the context of clustering, for being categorical variables in nature. As an example, '*Did a police officer attend the scene*' which is a 1 or 0 is excluded from the variable list, which would have little meaning or weight when a clustering algorithm is applied. Figure 1 shows a multivariate scatterplot between variable pairs with 1D kernel density estimate (KDE) of the marginal distributions on the principal diagonal. The data visualization of variables of interest is carried out using the Python data visualisation package: Seaborn [13].

The volume of the dataset is 1.26 GB which poses additional challenge in loading it in the CPU memory and cleaning for non-numeric entries. There are over two million entries in the datasets which is sufficiently large to train a machine learning model reliably. It would have been possible to create an even larger dataset by merging multiple of such similar datasets from further years and removing duplicates by matching their 'accident index' variable. However, for brevity the analysis has been focussed on one big dataset in [12] only.

B. Choice of Clustering as the k-means Algorithm

The *k*-means clustering is one of the simplest clustering methods using only simple geometric distance calculations (e.g. Euclidean distance or others). It has good scalability, and is capable of clustering large datasets in moderate to high dimensions at reduced computational expense. It is a far more efficient choice when compared to nonlinear methods like spectral clustering which uses a nearest neighbour graph to calculate distance between the clusters. However, this method would also be very computationally expensive to use for a very large dataset as done in this paper which has over one million entries due to increasing memory usage in the intermediate steps of the clustering algorithm. Hence, we restrict our study with only *k*-means clustering. In recent literature, few other studies applied computational statistics and machine learning methods for road accident analytics. These include *k*-means clustering [14], KDE maps of the severity index and categorization of hotspots [15], visual analytics and clustering for anomaly detection, unusual road behaviours, obtaining heatmap of accident prone areas and carry out hypothesis testing [16], as the most prominent options in terms of computational speed and robustness.

978-1-7281-5681-1/21/$31.00 ©2021 IEEE

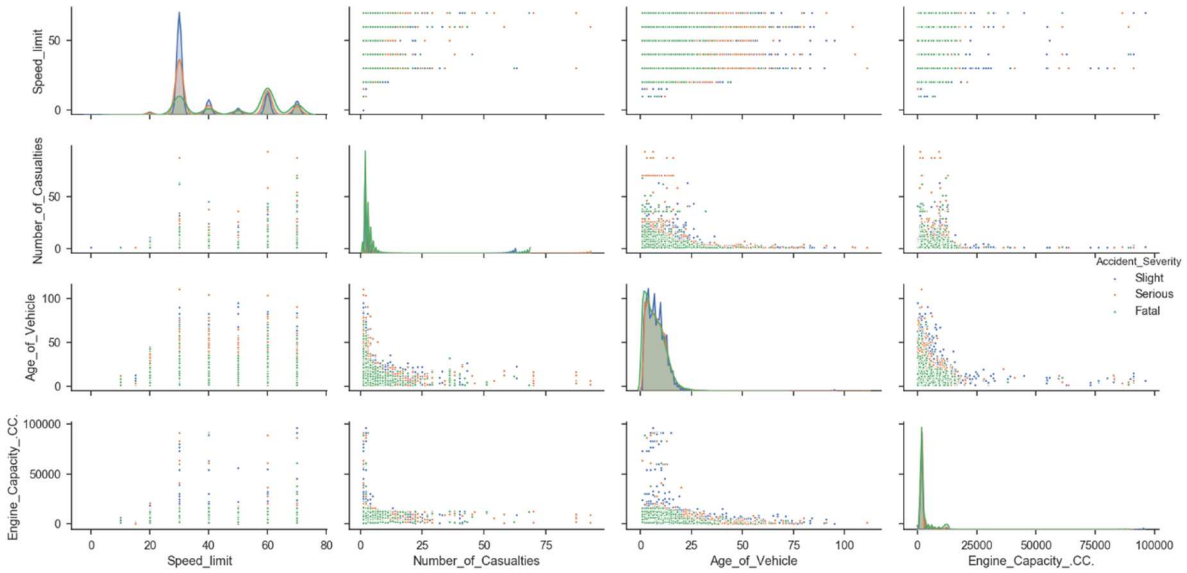

Figure 1: Multivariate scatterplot of the variables used in clustering.

C. *Hyperparameter Settings in the k-means Clustering*

The *k*-means clustering is a class of unsupervised machine learning algorithms that aims to separate *n* datapoints into *k* clusters. Unsupervised machine learning looks for patterns in a dataset with no prior labelling of the dataset, it uses a minimum of human supervision. Initially *k* centroids are randomly generated in the chosen multi-dimensional space. Every data point is then assigned to the closest centroid by minimum Euclidean distance criteria. Then the centroids are recalculated as the means of each new cluster and the assignment process are updated. Thus, the clusters gradually change with each iteration. The algorithm ends either when the centroids stabilize (no further change between two iterations) or when the given number of iterations have been performed. Due to the random nature of the initialization of the *k*-means clustering, it is not guaranteed to find the optimum arrangement. Due to this it is common to run the algorithm with several different initializations and then average the centroid coordinates that result. To calculate the optimum number of clusters for any given number of variables or features, one need to use the elbow curve, which matches the number of clusters against the mean square error (MSE) of all datapoints from their respective nearest cluster centroids. This process is repeated with increasing number of clusters but without plotting the points to save computing time. Using this method, we can determine the optimum number of clusters to use in mining this big dataset.

D. *Important Variables or Features in the Accident Data*

In order for variables to be usable in the *k*-means clustering algorithm they need to have a numeric value (float or integer) so that the Euclidean distance from the cluster centroids can be calculated. In this paper, we have used only variables that can be read as numerical variables for use within the *k*-means clustering. This includes 8 variables which are listed as an integer/float, or can easily be converted for example median value of the age bracket of the driver to feed as a single value to be used in the *k*-means algorithm as:

- latitude information,
- longitude information,
- number of casualties,
- number of vehicles,
- age of the driver,
- age of the vehicle,
- speed limit,
- engine capacity.

III. Results and Discussions

A. *General Overview of Road Accidents in the UK*

As a general context, here we provide a general overview of road accidents in UK, such as the density of accidents and where the urban or rural areas are in the UK. Figure 2 shows a decreasing trend of the number of road accidents in UK between 2005-2017 while it also shows small periodicity in the temporal pattern during winter due to fog and poor visibility. Figure 3 shows that most of the accidents are slight cases followed by serious and fatal cases. In Figure 4, we show the accident locations along with an overlayed 2D KDE plot which shows most of the accident-prone regions are around London and greater Manchester.

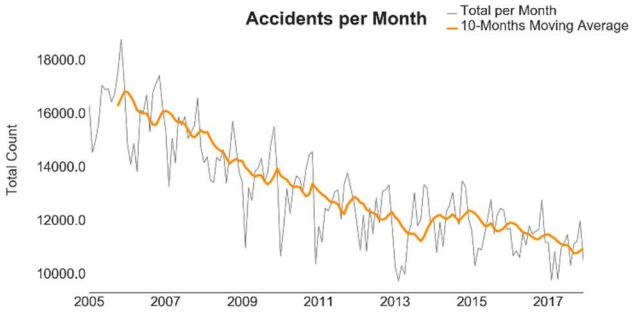

Figure 2: UK Accidents per Month between 2005-2017.

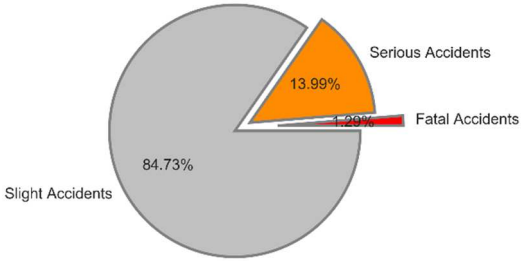

Figure 3: Pie-chart of the accident severity in percentages between the years 2005-2017.

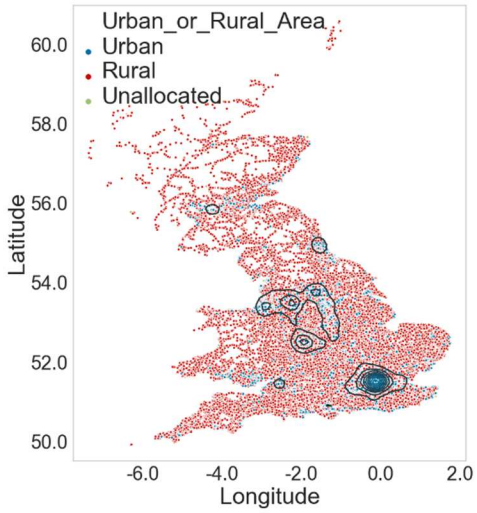

Figure 4: Scatterplot of UK accidents, split by 'Urban or Rural' with an overlayed 2D KDE plot.

B. *Higher Dimensional Analysis with Eight Variables*

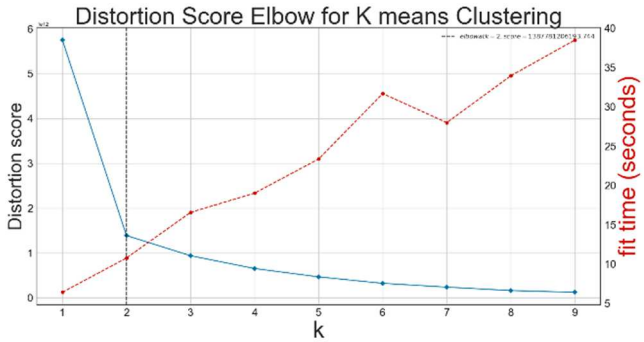

Figure 5: Distortion score elbow curve and computation time with increasing number of clusters for eight variables.

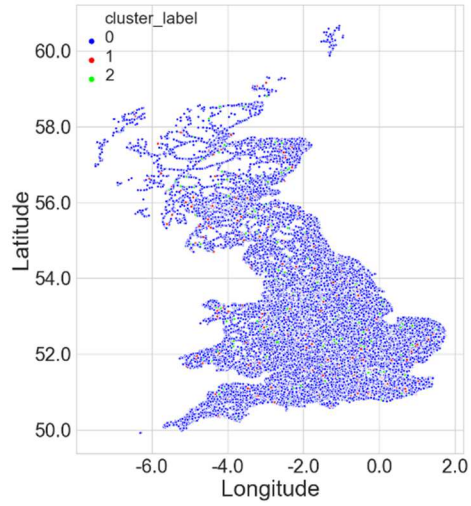

Figure 6: Scatterplots by latitude-longitude for three clusters obtained by the *k*-means clustering algorithm applied on all the eight variables.

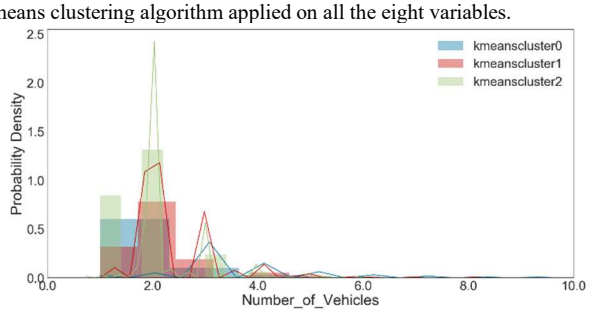

Figure 7: Univariate KDE plot of the number of vehicles split by clusters.

For our analysis, each datapoint has eight variables or dimensions viz. longitude, latitude, number of vehicles, number of casualties, speed limit, age of vehicle, engine capacity and age of driver. The datapoints are assigned to one of the three clusters. The three clusters were selected using the elbow method which identifies two clusters as optimal. However, the reduction in MSE vs. increase in the computing time for using three clusters is justifiable while also matching with the analysis in reduced dimensions. The datapoints in each of the 3 clusters can now be visualized in terms of scatterplots between the latitude-longitude as shown in Figure 6 or as univariate KDE plots of the respective variable viz. number of vehicles (Figure 7), number of casualties (Figure 8), speed limit (Figure 9), age of vehicle (Figure 10), engine capacity (Figure 11), driver's age (Figure 12).

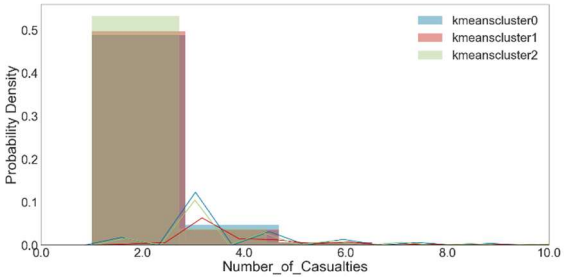

Figure 8: Univariate KDE plot of the number of casualties split by clusters.

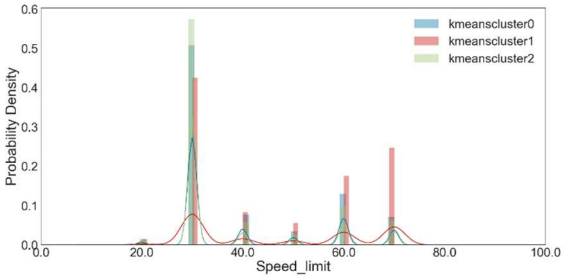

Figure 9: Univariate KDE plots of the speed limit split by clusters.

1) *Observation from the Geospatial Map for 8D Data*

As can be seen, cluster label 0 dominates the UK map in Figure 6, with highly focused spatial groupings of clusters 1 and 2 in narrow regions when the clustering is done on all 8 variables. This is also evident from the highly skewed number of datapoints in cluster 0 as shown in Figure 13.

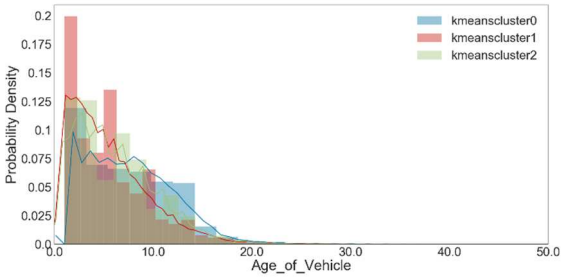

Figure 10: Univariate KDE plot of the age of vehicle split by clusters.

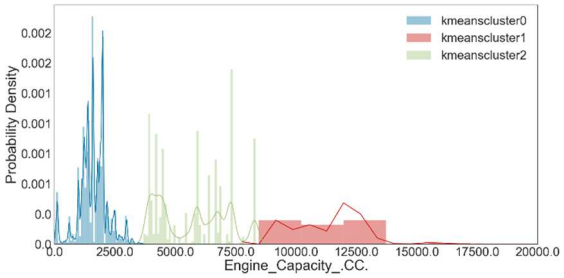

Figure 11: Univariate KDE plot of the engine capacity split by clusters.

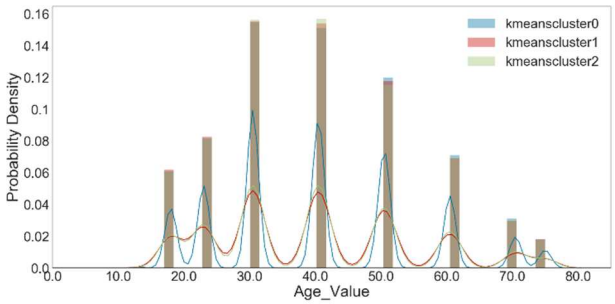

Figure 12: Univariate KDE of the driver's age value split by clusters.

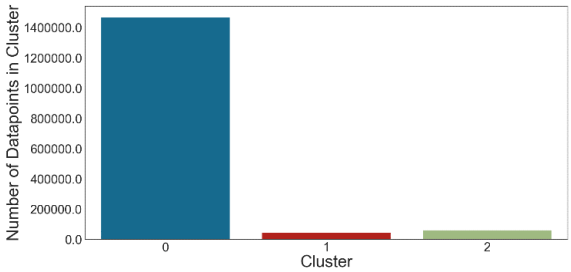

Figure 13: Bar graph of the number of datapoints in each cluster for 8D data.

2) Observation from Cluster 0

This cluster is clearly separated by the data points for engine capacity, in the range of 0cc to ~3000cc as observed from Figure 11. It is the largest cluster, showing that lower cc accidents are the most frequent in the UK. Moving forward from 10 years of age, these lower cc accidents have a higher probability density of being in older vehicles than in the other groups. Conversely this means that higher cc vehicles are more likely to be younger vehicles which fits the intuition.

3) Observation from Cluster 1

This cluster has engine capacity range of approximately 3000cc to 8000cc. The lower cc cluster dominates this clustering making it difficult to make observations relevant to other clusters. Cluster 1 is also slightly larger than cluster 2, this is due to higher cc vehicles being rarer.

4) Observation from Cluster 2

This cluster has engine capacity range of 8000cc and above. It has the highest probability density for the youngest vehicles from 0 years to approximately 4 years. This suggests that owners of these very high cc vehicles are most likely to have an accident within the first 4 years of owning the car.

5) General Observations

The probability density of the age of vehicle declines rapidly for all clusters. This could be used to suggest that owners of young vehicles are more likely to crash within the first few years of owning the car in combination with the fact that there are simply fewer older cars on the road due to malfunction or other reasons.

C. Reduced Dimensional Analysis with Seven Variables

As illustrated in the previous subsection, the engine capacity variable dominates the others showing clear separation of the clusters based on minimum overlapping intervals. In order to investigate the influence of other variables on the clustering, the engine capacity variable has been removed and the data is reduced to 7D. Repeating a similar elbow curve method to determine the optimum number of clusters, we observe from Figure 14 that three clusters is the optimal case for both computing time and minimum distance. Each of these clusters can now be visualized as 2D latitude-longitude geospatial map as shown in Figure 15 or the respective variable spaces similar to the previous section given in Figure 16-Figure 20. Using this reduced 7D clustering improves balancing the number of datapoints in each cluster as shown in Figure 21.

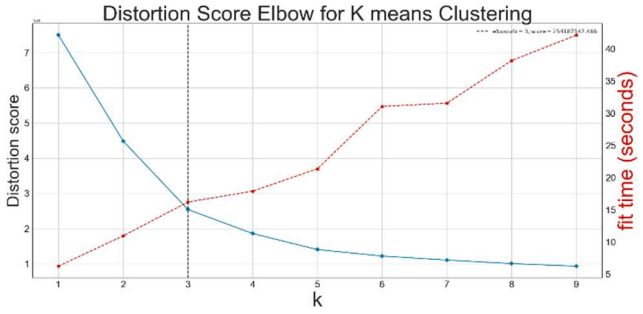

Figure 14: Distortion score elbow curve and computation time with increasing number of clusters for seven variables.

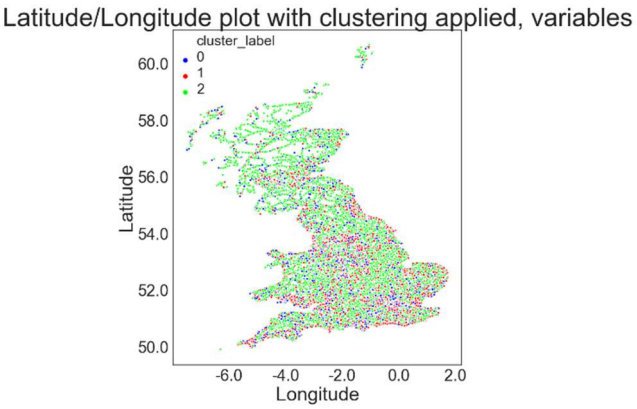

Figure 15: Scatterplots by latitude-longitude for three clusters obtained by the *k*-means clustering algorithm applied on reduced seven variables.

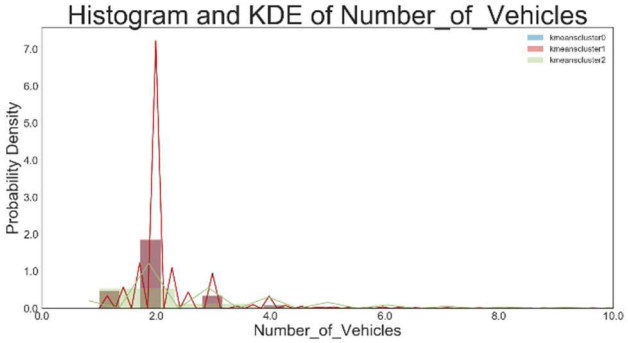

Figure 16: Univariate KDE plot of the number of vehicles for 7D data.

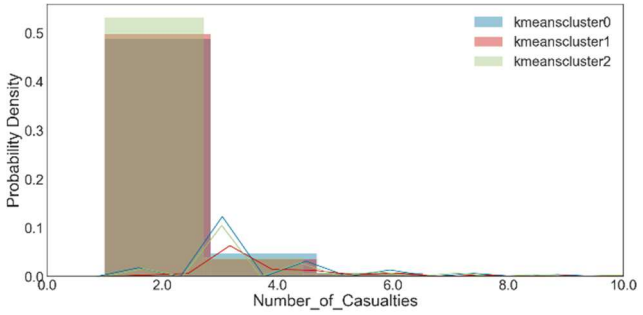

Figure 17: Univariate KDE plot of the number of casualties for 7D data.

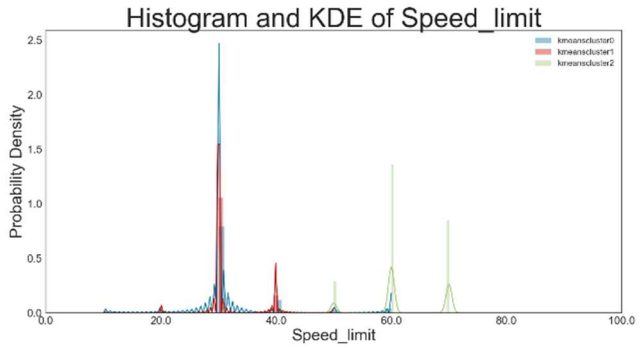

Figure 18: Univariate KDE plots of the speed limit split for 7D data.

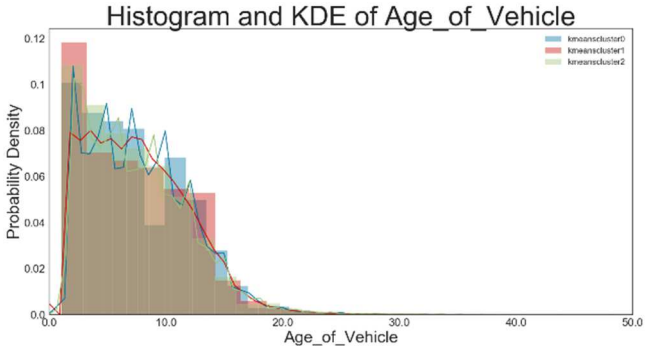

Figure 19: Univariate KDE plot of the age of vehicle for 7D data.

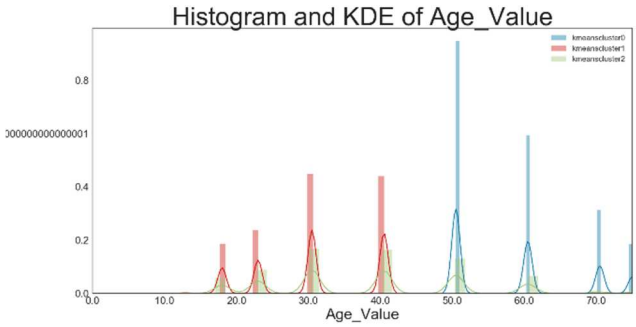

Figure 20: Univariate KDE of the driver's age value for 7D data.

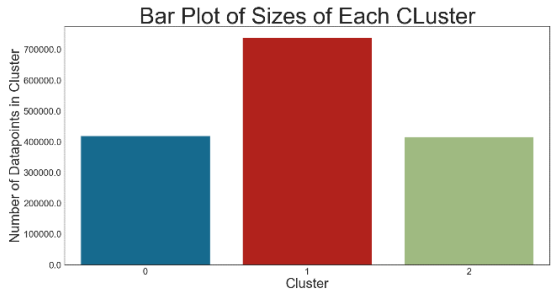

Figure 21: Bar graph of the number of datapoints in each cluster for 7D data.

1) Observation from the Geospatial Map for 7D Data

For the reduced 7D data, cluster 2 is found to be the most spatially widespread cluster, seemingly dominating rural areas with clusters 0 and cluster 1 being split between the urban regions. From the detailed plots in Figure 22-Figure 27, this appears to hold true, with what appears to be smaller towns and villages outside of Newcastle, London and Greater Manchester also split between clusters 0 and 1 in the reduced dimensional analysis.

2) Observation from Cluster 0

This is the largest cluster which contains accidents up to 40mph zones involving individuals in age brackets up to 40 years. It is almost exclusively positioned in urban areas, due to the speed limit influence on this cluster. This cluster potentially provides evidence to support the hypothesis that younger people driving in urban areas are the most likely to have a crash, interestingly it appears that when compared to cluster 1, they are also more likely to suffer a casualty as a result of the crash.

3) Observation from Cluster 1

This cluster contains accidents mostly in the 0 to 40mph range. As seen in Figure 18, there is a spike in probability density at the 60mph range though not nearly as much as cluster 2. This cluster contains accidents involving individuals in age bracket 50 and above. Interestingly, there is a lower number of casualties for this group. This perhaps suggests that accidents involving this older group are less likely to result in fatalities. Cluster 0 and cluster 1 overlap for number of vehicles, with a large spike at 2 vehicles far more than cluster 2. This may potentially support the hypothesis that urban crashes are more likely to involve multiple cars.

4) Observation from Cluster 2

Cluster 2 is differentiated from clusters 0 and cluster 1 by its much higher speed limit zones, involving accidents occurring in zones of 50mph to 70mph. This correlates with more rural areas under the national speed limit and motorways. It appears to have the highest probability density for multiple casualty accidents, there is a large spike at 3 casualties in Figure 17. Based on the histogram in Figure 16, this cluster also appears to contain more accidents occurring with only a single vehicle.

5) General Observations

Going against the hypothesis that older drivers would be driving older vehicles, there appears to have low correlation between age of the driver and age of the vehicle. This is because all three clusters follow a similar path. There are a few spikes in the probability density throughout. But all three clusters look very similar in terms of the age of the vehicles.

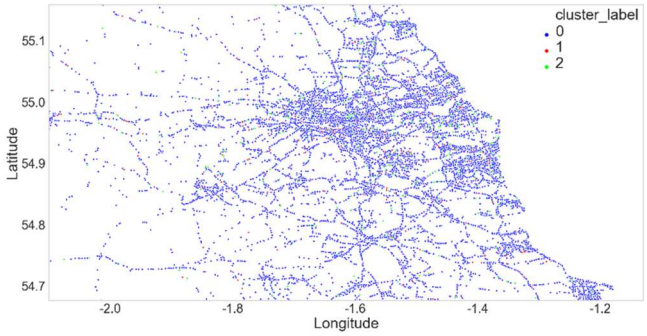

Figure 22: Newcastle localized geospatial map of higher dimensional (8D) cluster analysis. An enlarged version of Figure 6 around Newcastle.

D. Localised Latitude and Longitude Plots

Considering the macroscale clustering of Figure 6 and Figure 15, further enlarged versions of these geospatial maps are now generated for a more localized analysis around Newcastle, Greater London and Greater Manchester and Liverpool which is highlighted as the high density regions for accident in Figure 4 before the clustering was carried out. This is to show these high-density accident regions in greater detail in order to see the finer structures of the clustering. Figure 22, Figure 24, Figure 26 show the 8D clustering analysis applied to Newcastle, London and Manchester respectively. The 8D analysis shows that the macro-scale geospatial maps have little spatial correlation between the clusters, which are primarily being split by engine capacity.

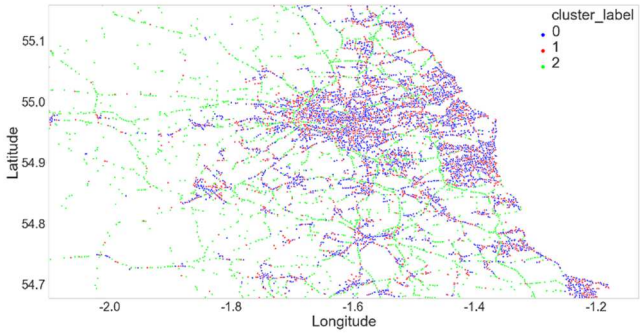

Figure 23: Newcastle localized geospatial map of higher dimensional (7D) cluster analysis. An enlarged version of Figure 6 around Newcastle.

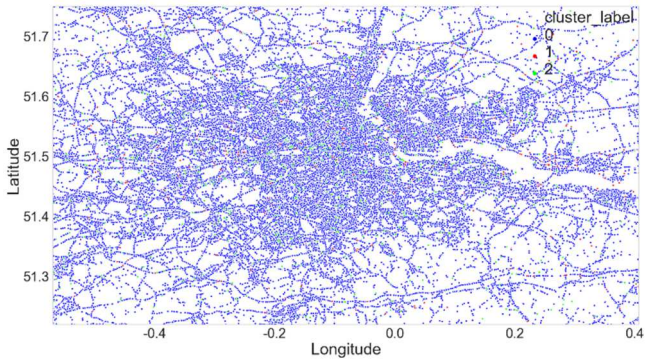

Figure 24: Greater London geospatial map of higher dimensional (8D) cluster analysis. An enlarged version of Figure 6 around London.

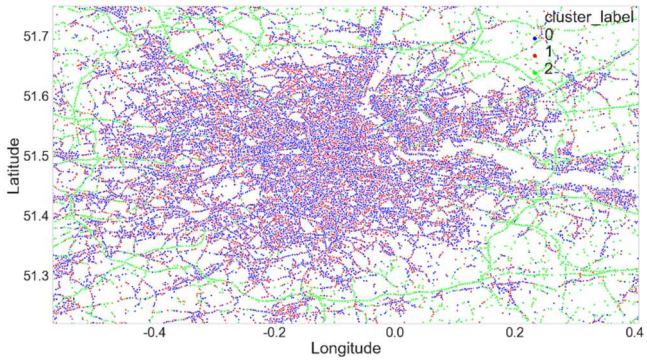

Figure 25: Greater London localized geospatial map of higher dimensional (7D) cluster analysis. An enlarged version of Figure 6 around London.

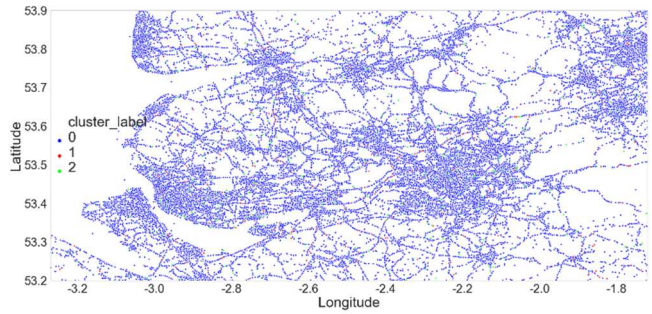

Figure 26: Greater Manchester and Liverpool localized geospatial map of higher dimensional (8D) cluster analysis. An enlarged version of Figure 6 around Manchester and Liverpool.

Figure 23, Figure 25, Figure 27 show the reduced 7D analysis around Newcastle, London and Manchester regions which reveals that clusters 0 and 1 not only apply to urban cities but also smaller towns and villages. This is because a heavily influential feature in the reduced 7D analysis is the speed limit. These enlarged maps of the reduced 7D analysis show that motorways are considered as a part of cluster 2. This will be due to speed limit. However, it is interesting to note that truly rural crashes correlate enough with motorway crashes which need to be considered a part of the same group. Here again motorways passing around city centers are visible as being a part of cluster 2. This suggests that a more in-depth breakdown of road class and its correlation with the clusters formed may be interesting to investigate further. By looking at the enlarged pictures we can observe the finer structures of the clusters and the relationships between the variables which do not change with the geospatial scale of the observation.

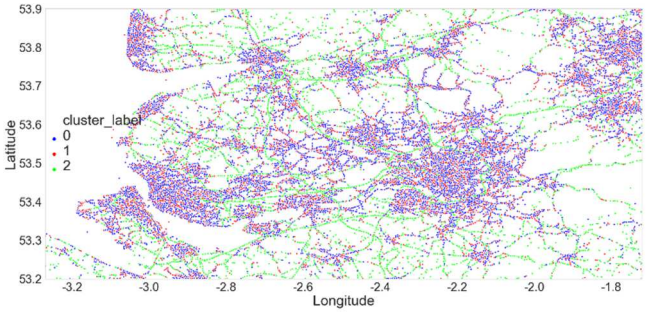

Figure 27: Greater Manchester and Liverpool localized geospatial map of higher dimensional (7D) cluster analysis. An enlarged version of Figure 6 around Manchester and Liverpool.

In Figure 28 and Figure 29 we show the latitude and longitude plots of Motorway class and A class roads. This excludes A class roads in urban areas which were all included in cluster 2 of the reduced dimensional analysis. The details can be seen in the localised plots in Figure 25-Figure 27 which include motorways. There appears to be no spatial correlation at least in these variables, though the connection may be worth further investigation.

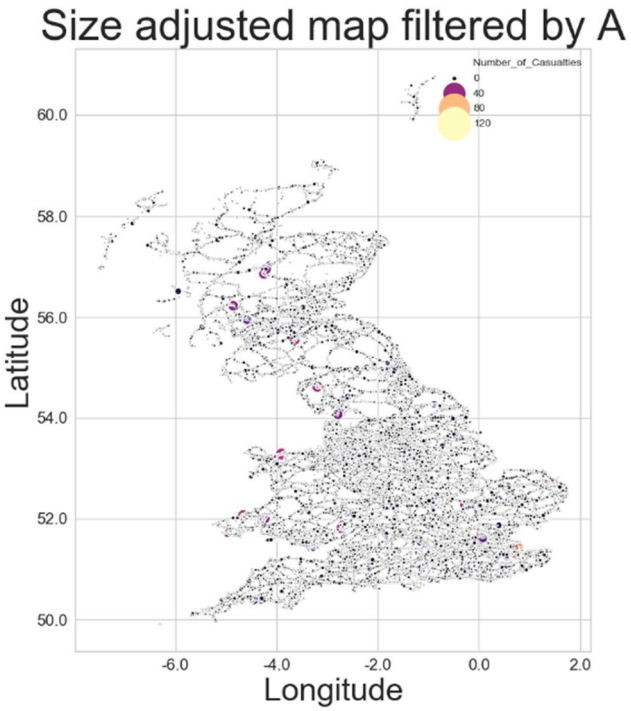

Figure 28: Latitude longitude plot of A class roads coloured and sized by the number of casualties.

IV. Future Scope of Work

The future scope of work would be to look in detail at other clustering and statistical modelling methods for a detailed comparison of additional information gained vs. complexity

of the models. A greater number of variables can also be included in the analysis after researching and adapting each one to fit the basic *k*-means clustering inputs, such as correlating weather conditions from local or national weather station e.g. temperature or wind condition data with accidents in the UK. These may provide additional insights into what weather conditions cause what kind of accident and if any specific group is more at risk. It would also be interesting to look in further detail at what role road class plays in road accidents, given that motorways and rural crashes are similar enough to be put into the same category e.g. high speed, more likely to be a single vehicle leading to greater chance of a fatality occurring. This encourages further study into other road classes and how far we can classify, or cluster vehicular accidents based on the class of roads, the accident occurs on.

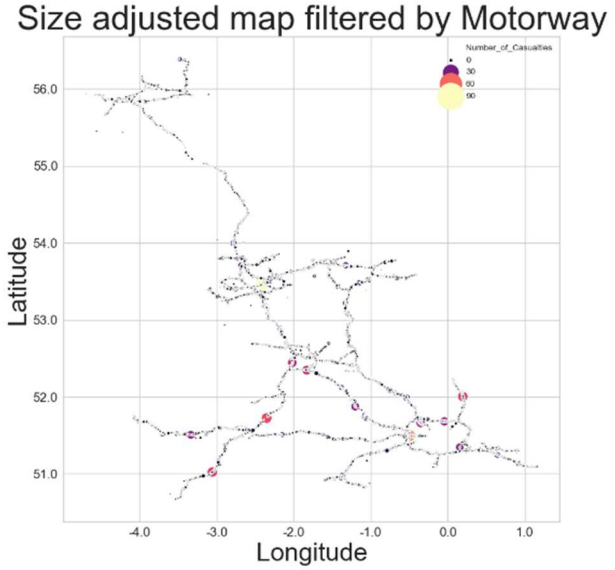

Figure 29: Latitude longitude plot of Motorway class roads coloured and sized by the number of casualties.

V. Conclusion

The *k*-means has been shown to be a good choice for accident analytics involving big open dataset for UK. However, other clustering methods could have been explored depending on the computing power requirements for handling such big data. The *k*-means combined with a breakdown of each cluster using KDE gives a lot of information about the variables under investigation. A simple use of *k*-means clustering on the UK road accident dataset supports boosting evidence for the hypothesis about vehicular accidents such as '*Young people are more likely to suffer an accident*' as well as provided more detail beyond this such as '*Young people driving in urban areas are more likely to suffer an accident*'. The dataset analysed here can be expanded in future. Further investigation could be carried out to explore the changing influence of vehicle accident variables over time, and which variables were more prominent in the past and what will become prominent in the future. The methodology used in this paper can easily be applied to other datasets, or datasets covering more years in the UK. It is computationally less expensive to use simpler clustering methods which provides a lot of insight into variables of accident analytics and prevention and allied topics in general. It may also be worth carrying out similar data analytics combining datasets from other countries. It may provide additional information that can be used in further refining the machine learning models to predict the probability of an accident for an individual, providing additional information for studies on driverless cars.

Acknowledgment

Christopher Sinclair thanks the EPSRC Summer Vacation Internship from the CEMPS, UoE for supporting him.

References

[1] D. D. Clarke, P. Ward, C. Bartle, and W. Truman, "Killer crashes: fatal road traffic accidents in the UK," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 764–770, 2010.

[2] C. Wang, M. A. Quddus, and S. G. Ison, "Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England," *Accident Analysis & Prevention*, vol. 41, no. 4, pp. 798–808, 2009.

[3] A. Jain, G. Ahuja, D. Mehrotra, and others, "Data mining approach to analyse the road accidents in India," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2016, pp. 175–179.

[4] A. V. Sakhare and P. S. Kasbe, "A review on road accident data analysis using data mining techniques," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–5.

[5] L. Li, S. Shrestha, and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2017, pp. 363–370.

[6] M. G. Mohamed, N. Saunier, L. F. Miranda-Moreno, and S. V. Ukkusuri, "A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada," *Safety Science*, vol. 54, pp. 27–37, 2013.

[7] J. De Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks," *Accident Analysis & Prevention*, vol. 51, pp. 1–10, 2013.

[8] V. Prasannakumar, H. Vijith, R. Charutha, and N. Geetha, "Spatio-temporal clustering of road accidents: GIS based analysis and assessment," *Procedia-Social and Behavioral Sciences*, vol. 21, pp. 317–325, 2011.

[9] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[10] M. A. Mondal and Z. Rehena, "Identifying Traffic Congestion Pattern using K-means Clustering Technique," in *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, 2019, pp. 1–5.

[11] M. Y. Choong, L. Angeline, R. K. Y. Chin, K. B. Yeo, and K. T. K. Teo, "Modeling of Vehicle Trajectory using K-Means and Fuzzy C-Means Clustering," in *2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 2018, pp. 1–6.

[12] "UK Road Safety: Traffic Accidents and Vehicles." [Online]. Available: https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles

[13] E. Bisong, "Matplotlib and Seaborn," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Springer, 2019, pp. 151–165.

[14] P. A. Nandurge and N. V. Dharwadkar, "Analyzing road accident data using machine learning paradigms," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 2017, pp. 604–610.

[15] K. G. Le, P. Liu, and L.-T. Lin, "Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi, Vietnam," *Geo-spatial Information Science*, vol. 23, no. 2, pp. 153–164, 2020.

[16] I. Kalamaras, A. Zamichos, A. Salamanis, A. Drosou, D. D. Kehagias, G. Margaritis, S. Papadopoulos, and D. Tzovaras, "An interactive visual analytics platform for smart intelligent transportation systems management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 487–496, 2017.